Content Summarization using BERT

By
Aditya Tushar Wadnerkar
Rimzim Thube
Shree Gowri Radhakrishna
Vijaylaxmi Nagurkar

Abstract

In the age of information we have seen an explosion of data in the content generated by end users on various platforms over the internet. In order to get information from a source we need to go across hundreds of files be it in the form of text articles or illustrations using videos on sites like youtube. It becomes cumbersome to go through all the information available on the internet based on our advanced search to identify the best article we are looking for. Hence this project helps users **summarize the information** over the internet conveyed by the **data source**; be it a **research paper, blog,** or an illustration **video on youtube**. We quickly use the **narration feature on variable speeds** as per convenience to save the user's time understanding the context of the source.

Proposed Study

We will be using one of the latest state of art mechanisms called BERT, for generating a meaningful summary of any content provided on the internet for faster interpretation of each source. It will be based on a RESTful service which would utilize BERT model for text embeddings and K-means clustering for identifying the sentences which are the closest to centroid for selection of summary. This model will be used for the extractive text summarization.

Data Source

This project will use the free available coronavirus dataset having information about the number of cases, number of deaths and number of recovered people. It will have additional information about countries in which the cases have occurred.

https://data.humdata.org/dataset/novel-coronavirus-2019-ncov-cases

Input

The input will be the transcript extracted from a text/pdf/word/audio/video file or the required content which would be scraped form an input source link from the internet.

Output

The project will have the following outcomes:

1. It will print the summary of the source which could be a textbox/pdf/word/audio/video file.
2. It will take the youtube video link, blog post link or a link to a research journal as the input and give a summary of what is discussed in the content.
3. It will have an option to narrate the summary generated by the application on variable speeds to save the user's time.

Roles and Responsibilities

Literature review	Rimzim/Vijaylaxmi
Data Preprocessing	Gowri/Aditya
Architecture study	Rimzim/Vijaylaxmi
Model training	Gowri/Aditya
Visualization	Team